University of Alberta

Virtual Reality and McDermott's Model of the Mind

by

Jim Gordon Stenberg

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Master of Arts

Department of Philosophy

Edmonton, Alberta

Fall 2005

0-494-09068-5

Library and Archives Canada

Published Heritage Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et Archives Canada

Direction du Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file *Votre référence*
ISBN:
Our file *Notre référence*
ISBN:

NOTICE:
The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:
L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

Abstract

I examine the computational model of the mind Drew McDermott presents in *Mind and Mechanism* and supplement it with the explanation of virtual reality David Deutsch gives in *Fabric of Reality*. The resulting model gives a non-eliminativist account of qualia and also contributes to a better understanding of free will. Although McDermott professes functionalism I argue that the notion of reference employed in his model cannot be accounted for in functionalist terms.

Table of Contents

Introduction 1

Chapter One
Syntax, Semantics and Self-Reference

1.1	The Objectivity of Syntax	4
1.2	The Objectivity of Bits	9
1.3	Semantics	12
1.4	The Self-Model	17

Chapter Two
Illusions and Virtual Reality

2.1	Different Kinds of Illusion	21
2.2	The Virtual Reality Relation	23
2.3	Real Ideals	26
2.4	Functionalism and Antifunctionalism	29
2.5	The Self-Model in terms of VR	32
2.5.1	Sense Perception as VR	34
2.5.2	Internal Experiences	35
2.5.3	Why A and B shouldn't be distinguished	37

Chapter Three
Qualia

3.1	What are Qualia?	40
3.2	Chalmers and McDermott	45
3.3	Other Perspectives on Qualia	49
3.4	Contentious Qualia	51

Chapter Four
Free Will

4.1	McDermott's Theory	55
4.2	The Robot and the Bomb	59
4.3	Supplementing McDermott's Argument	60
4.4	Refuting McDermott's Argument	62
4.5	An Alternative Argument	64
4.6	More about Free Will	69

Conclusion 72

Literature Cited 73

Introduction

In this thesis I examine Drew McDermott's explanation of mind in *Mind and Mechanism*, and connect this explanation to some broader issues regarding functionalism and the reality of mental states. Although *Mind and Mechanism* was written partially as a response to Chalmers' *The Conscious Mind* I will focus on what McDermott says, not what he is replying to. In discussing McDermott I will concentrate on his treatment of phenomenal consciousness and free will. His treatment is not perfect; where I have found shortcomings in his theory I have endeavored to present a supplementary or alternative line of argument that leads to the same general conclusions.

McDermott's approach to the mind is based on the idea that the brain is a computer, and that it maintains a computational model of itself; the self-model. In McDermott's theory both qualia and the free will are direct results of the limits on how a system can model itself. Qualia and free will derive from imperfections and limitations of the self-model, and are not "something extra" that a conscious organism possesses. A large part of this paper will consist of presenting the details of this theory, which McDermott refers to as a theory of "virtual consciousness."

The theory of virtual consciousness is a computational, quasi-eliminativist theory of qualia and free will. It basically states that qualia arise when a

computational system distinguishes between an object's being a certain way and only appearing that way, and that free will arises when a system can change a decision based on the consequences it foresees for that decision. This is "virtual consciousness" and his argument is that what we call consciousness is nothing more than this. The qualitative aspects of perception and decision making are presented as essentially illusory.

Unpacking McDermott's theory begins with the requirement that the brain can be a computer in an objective and significant way. If the brain is a computer only in a technical, trivial sense, or if a computer only exists relative to a conscious user, then his theory will not work. These objections are associated with John Searle and Hilary Putnam, and a significant part of McDermott's book consists of a detailed response to these objections. In particular McDermott tries to show that syntax is grounded in physics. I will consider this response (on the whole a convincing one) before presenting McDermott's theory of virtual consciousness.

McDermott's theory is based on the notion of a self-model. Being a self-model is a semantic notion,[1] and so McDermott has to show that there is an objective way in which a computational model's semantics can be grounded in its syntax. This task is just as fundamental as showing that syntax is objectively grounded in physics, but McDermott gives a less detailed account of it. There is a lot of hand-waving and appeals to the objectivity of scientific explanation. Although it is less detailed, I don't think McDermott's treatment of semantics is unconvincing. But there is one point that should be noted: while McDermott's definition of a computer is compatible

[1] i.e. to determine if a symbol "I" refers to the system that generates that symbol, you need to have a theory of how symbols refer to their objects; you need an account of semantics.

with functionalism, the notion of being a self-model is not. There can be two functionally equivalent systems (and with the same causal powers, no less) only one of which is a self-model. In other words, self-reference is not a functional notion.

In an attempt to be sensitive to the controversy over the reality of mental phenomena, I will draw on material from David Deutsch's *The Fabric of Reality* and Jack Copeland's *Artificial Intelligence*, material which has to do with illusions, appearances and simulations. I will extend McDermott's argument that the brain is a computer and say that it is a very special kind of computer- a computer which generates a virtual reality simulation. The reality being simulated is the world as a whole, and the self-model is that part of the simulation that models the system which performs the simulation. The computational objects in the simulation have an objective reality; this follows from the objectivity of the computer's syntax. Being a simulation, an illusion of sorts, is thus not incompatible with being real.

In many ways the treatment of free will is analogous to the treatment of qualia. McDermott regards free will as being an illusion, a side effect from the way that a system models itself. But just as the treatment of qualia does not lead to an eliminativist position, neither does the treatment of free will. I believe that a sensitive treatment of what it means to be part of a virtual reality simulation can do justice to the insights of the functionalist position without denying the reality (properly understood) of phenomenal consciousness and the experience of free will.

Chapter One

Syntax, Semantics and Self-Reference

1.1 The Objectivity of Syntax

McDermott has to define a computer in such a way that the concept is not trivial. For if every macroscopic object meets the definition of a computer, the definition is clearly too loose. The accusation of triviality is associated with John Searle (1992) and Hilary Putnam (1988), who have argued that the multiple realizability of computers entails universal realizability, and that something's being a computer is therefore vacuous. Searle famously said that

> ...the wall behind my back is right now implementing the Wordstar program, because there is some pattern of molecule movements that is isomorphic with the formal structure of Wordstar. (*Rediscovery of the Mind*, pp. 208-209.)

Putnam's conclusions are similar to Searle's "universal realizability" result. Putnam shows that "[e]very ordinary open system is a realization of every abstract finite automaton" (Putnam 1988, p. 121). The only requirement is that the system have enough distinguishable parts; a rock would do. The proof involves defining a correspondence between molecular positions and syntactic units in a very

gerrymandered and disjunctive (but formally acceptable) manner. Since something's being a computer is a syntactic notion, it is not surprising that peculiar syntactic definitions might lead to peculiar results.

The most obvious response to Searle and Putnam is that a rock or wall could not be a computer because there is no practical way to connect a rock or wall to a keyboard, monitor, or printer in a way that would be useful to a human user. It would take very special equipment and an extraordinary intelligence to perceive the motion of molecules in a wall to implementing a Wordstar program. Searle would accept the notion that computers can only be talked about in terms of a human user. He believes that "whenever we use a computer, we are the ones who attribute the syntax to the machine and interpret its actions as computational." (*Rediscovery of the Mind*, p. 214) This response is not available to McDermott because that would imply that something's being a computer depends on a conscious user, and as he puts it,

> If my brain's being a computer turns out to explain consciousness, it had better not be necessary to appeal to the judgments of conscious entities to explain what a computer is. (p. 167)

McDermott wants to explain consciousness in terms of computation in a non-vacuous and non-question begging way. His definition of computer therefore needs to exclude both Searle's wall and Putnam's rock from being a computer, and must ensure that being a computer is an objective property. McDermott's definition (and explanation of terms) spans 3 pages of *Mind and Mechanism*, but is straightforward to summarize[2]:

---

[2] Much of the following two paragraphs is lifted directly from McDermott, but I have rearranged and edited some of what he says. These are some of the changes I make. McDermott uses *S* to refer both to

A computer is a physical system whose outputs are a function of its inputs, where these concepts are defined as follows. The system must have a set of *states*, although there is no requirement that they be discrete states. Most of the time, when McDermott talks about system states, he means partial states, so that a system can be in several states at once. (See p. 169)

Input and output *decodings* are mappings from certain partial states to range and domain sets. Given any (partial) state S of this system, the input decoding $C_I(S)$ is either a particular number, or $\perp$ if the input is not determined in state S, and similarly for the output decoding $C_O(S)$. The system needs to have a distinguished state class A such that whenever the system is in A the output class it is in can be determined. An *input/output* decoding of a system is just the ordered triple $<C_I, C_O, A>$. Using this notation, a computer is defined as follows. The system M computes a function F with respect to a decoding $<C_I, C_O, A>$ if and only if whenever it is in a state S with $C_I(S) = x$ that causes it at some observable later time to be in a state S' with $C_O(S') = F(x)$. (See p. 170)

The notion of causality employed in this definition supports counterfactuals. However the system gets into S, it ends up in an output class in the state S', and if it were not in S, it would behave differently thereafter and might not end up in S'. The systems considered by Searle and Putnam do not share this property.[3] McDermott

---

the system and the state the system is in, which is confusing; I refer to the system as M. He omits "class" from the initial definition of A, but then describes it as a class of states in a subsequent paragraph; I modify the definition accordingly. McDermott also has a stray "it" included in his definition of a computer, so that part of it reads "...that causes it at some observable later time it to be in a state..." Finally, he writes S instead of S' in the equation $C_O(S') = F(x)$. This rewritten version is, I think, what he means to say, but it would be cumbersome to indicate exactly which words are his, and which are mine. The reader is referred to the indicated pages of *Mind and Mechanism*.

[3] For a thorough discussion of Putnam and Searle's objections, see David Chalmers' "Does a Rock Implement Every Finite-State Automaton?"

also suggests a "continuity requirement" so that the system will not be too sensitive to small perturbations; both Searle's wall and Putnam's rock are extremely sensitive to environmental changes, so much so that providing input would require completely changing the gerrymandered definitions which allegedly establish them as computers. A computer that cannot accept input without being destroyed is clearly not very much of a computer.

There is some truth in what Searle and Putnam say. It is true that any physical system can be interpreted in different ways, and can be viewed as implementing several different computations. For example, it is trivially the case that each physical object, even one that does nothing, can be interpreted as implementing a 1-state computer. And ordinary desktop computers can usually be interpreted as implementing multiple programs at a time for the simple reason that they *are* usually implementing multiple programs at one time. But there are very few decodings under which a system is a computer, and so it is not the case that every physical system simultaneously implements every program. The notion of being a computer is thus not vacuous.

McDermott's definition of a computer shows how much the definition of computer has shifted over time. Just as a 'typewriter' originally referred to the person doing the typing and not the machine that was being typed on, so a 'computer' originally referred to the person who calculated ('computed') something. McDermott's definition accords more with modern sensibilities, which considers a computer to be a machine, not the machine's human user. Searle's insistence that machines don't compute anything is more intelligible when the old-fashioned

definition of computer is recalled. In the old sense, existing machines are not computers. Searle's insistence that computation involves consciously following rules (like a human computer would), not merely engaging in behavior that can be described by rules (as a machine computer would) also traces, I think, to the old meaning of the word. A lot of the difference between Searle and McDermott can be explained by the fact that they use the same words in different ways.

So much for the first part of McDermott's task, that of finding a non-trivial definition of computer. How about the alleged observer-relativity of a computer? None of the concepts invoked in McDermott's definition of a computer are observer relative, not even the decoding. As McDermott explains,

> ...whether or not something is a computer relative to a particular decoding is a matter of objective fact. It's no more subjective than the speed of an object, which cannot be measured without a frame of reference. It doesn't matter whether the decoding is of any use to anyone, or has even occurred to anyone. (p. 171)

I am not sure this addresses Searle's observation that computation is user relative. Searle says that we are the ones that attribute syntax to a computer, since syntax is not intrinsic to physics. Although McDermott says "input/output decoding" instead of "syntax," it seems as if he concedes this point, for he admits that although

> One might wish for a "natural" decoding that would stand out as the "obvious" one by which to evaluate the claim that a system is a computer... [but] one need not choose any decoding, elegant or not, as the "correct" one. (p. 172)

McDermott says that there is no natural decoding of an object, and that it is up to us to choose a decoding. This seems equivalent to Searle's claim that the physics of an object does not determine its syntax, and it is up to us to choose which syntax to ascribe to it. Searle admits that there are features of an object which facilitate one ascription rather than another, (perhaps he should say that physics does not *completely* determine syntax), but this does not undercut his point about user relativity, a point that does not seem to be adequately addressed by McDermott.

1.2 The Objectivity of Bits

On the other hand, note that to establish the fact that digital computers are computers McDermott needs only a very limited result; that there is some objective sense that a digital computer manipulates bits. And this seems straightforward.

What goes on in a digital computer can be explained in terms of voltages. That is to say, the voltages in the various areas can be expressed as real numbers, and a physical story told about how a particular pattern of voltages leads to a different pattern of voltages. The structure of the transformation of patterns is determined by physical laws and the boundary conditions of the physical system. This is the physics approach to the system.

The syntax of the system derives from the fact that in a digital computer the physical story can be expressed in a far more concise form by expressing the patterns of voltages as patterns of bits. "0" corresponds to a certain range of voltages, and "1" corresponds to another (non-overlapping) range of voltages, and various logical operations correspond to the physical processes in the computer whereby the

transformation of the patterns of voltages is given a structure. Aside from trivial variations (relabeling 0s as 1s and vice versa, or expressing everything in hexadecimal), there is a unique, simplest, structure preserving mapping from physics to symbols: no simpler mapping preserves the structure, and any other structure preserving mapping is more complex, often much more complex. Consider how much more information (and complexity) there is in a real number (expressing a voltage) than in a binary digit.

The binary digits and logical operations are the syntax, and they are based on the physics in a way that seems far from arbitrary. The fact that so much information (about the transformation of voltages) can be described so concisely is remarkable, and it requires very careful engineering to ensure that a binary decoding works so well when applied to the voltages inside the computer.

One could, of course, apply other decodings to a digital computer. For example you could choose the two-state decoding defined by C_1 = (plugged in = 0, not plugged in = 1) and C_0 = (humming = 0, not humming = 1). The computer would then compute the identity function (i.e. f(0) = 0 and f(1) = 1) if and only if it hums when it is plugged in, and doesn't when it is not. While it is true, I suppose, that choosing one decoding over another is technically arbitrary, it seems very silly, in the case of digital computers, to choose this second decoding over the first.

It should also be pointed out[4] that the typical user of a digital computer does not interpret the voltages in the various areas of the computer; without special instruments these voltages are not even perceptible. Instead the user interprets what

[4] Compare Domsky p. 35, footnote 66.

he sees on the computer screen. A user sees text instead of squiggles when using a word processor, for example.

One could regard this as a sign that there is a social division of labor[5]; perhaps a loose community of computer engineers and machine language coders is the entity who interprets the physics of a digital computer and assigns a particular syntax. But this is not exactly correct either. If every engineer, programmer and machine language coder were to die suddenly, the user would still see text instead of squiggles when using a word processor. The case of the ignorant user shows that it is untrue to say that it is the user who ascribes the syntax to the computer's inner workings.

It is more accurate to say that to explain the user-level functioning of computers, as exemplified by word processors, we would need to invoke programming syntax and mechanical syntax. An explanation phrased solely in terms of the voltages inside the machine would not be illuminating. An illuminating explanation would require, at the very least, the ability to speak in terms of bits and logical operations, and would very likely require reference to various machine languages and higher level programming languages.

Even if it were the user who ascribed the syntax (aided by very superior knowledge and the use of special voltage detectors), the mantra of "user relativity" would not explain why the user ascribes the syntax of a word processor to the computer rather than to the nearest wall. Although the latter task requires superior knowledge and special equipment, so does the former. This suggests a possible response to those who, like Searle, would claim that being a computer is arbitrary, and that a section of wall computes the Wordstar program just as truly as a digital

[5] Domsky refers to Wesley Cooper as the origin of this suggestion. Domsky p. 35, footnote 66.

computer does. Say that yes, being a computer is just as arbitrary as the choice between using a computer or using a wall as a word processor; in other words, not arbitrary at all.

McDermott's claim that computing is objective seems reasonable. That is, there is an objective sense in which bits have an objective reality, and are not just interpretations made by a conscious mind of what the different voltages in a machine might be.

1.3 Semantics

Some terms in computer science straddle the border between syntax and semantics. Although McDermott is careful (as we shall see) to define a computational model in a way that does not imply that the model refers to anything, it is natural to say that a model is a description of whatever is being modeled. Computational objects inside a model are even more difficult to describe without using semantic notions. As a general term, a computational object "encapsulates data and processing... [and] provides a set of capabilities that can be used by other computational objects."[6]

Just as McDermott needs an objective syntax to ensure that a brain counts as a computer, he needs an objective semantics to ensure that symbols in the brain/computer denote anything. This is because McDermott's theory of consciousness "rests on the claim that conscious brains manipulate models of themselves- that is, symbols denoting aspects of themselves." (p. 196) If the

[6] This definition can be found by googling "encapsulates data and processing". It seems to be present on unrelated websites. One copy of the definition can be found directly at www.cordis.lu/infowin/acts/rus/projects/screen/glossary/glossary.htm .

explanation of consciousness depends on what symbols denote, it had better not be necessary to appeal to the judgments of conscious entities to determine this. McDermott's definition of computer appealed to a particular decoding to determine what the syntax was. Unfortunately "[t]here's nothing corresponding to a decoding that just settles by convention the question of what symbols mean." (p. 210)

McDermott rejects the notion of arbitrarily assigning an interpretation to the syntax, for that would show only what the syntax could mean, not what it does mean (p. 209). To determine what the syntax means, McDermott appeals to the notion of a "most harmonious semantics." He argues that it is a scientific question what the most harmonious semantics is for a given system of symbols. He explains that "a semantics is "harmonious" if it provides a coherent story about the relationship between symbols and sensorimotor events." (p. 202)

"Coherent story" is meant in the sense that scientific theories and hypotheses provide coherent stories about the phenomena they describe. If science is objective, reasons McDermott, then so are claims about semantics (and specifically, claims about what symbols denote). There is room, of course, for alternative theories about what a given symbol might refer to. McDermott gives the example of a fly coming near a frog:

> When we posit a symbol to denote the presence of a fly, we may suppose that the symbol means "insect present." But some other analyst might express the meaning as "edible object present." There is no way to settle the issue, but also no particular reason to settle it. In the frog's environment, insects and edible objects are essentially the

same thing. The frog will make a certain number of errors (when it misclassifies something as an insect when it isn't, or vice versa), but as long as the number is small, the frog will prosper. (p. 197)

It is important to note that the kind of "objective semantics" that McDermott needs is a rather simple one. It is not a full-fledged theory of intentionality. What McDermott is looking for is a detailed computational explanation of how a frog's brain enables the frog to catch flies. Such an explanation must include a description of how the frog's brain represents a fly when it is seen; this representation involves both syntax and semantics.

It is also important to note that the methodology of science involves taking the "third person perspective." The reports, actions and judgments of the system have to be explained by the coherent story, but the story is for an observer outside the system being explained.

McDermott's discussion of the vestibular-ocular reflex (VOR) is another illustration of the kind of explanation he is interested in. The VOR involves signals from the inner ear being used to keep the eyes oriented toward a fixed target even when the head moves (p. 176). Searle argues ("The Explanation of Cognition" p. 123) that the VOR should be understood not as a computation but in purely biological terms. McDermott replies that expressing the VOR as a computation allows generalizations that cannot be rephrased in the terms of the underlying biological medium:

> *The VOR evolved because the information it provides to the visual system is valuable.* [emphasis McDermott's] This generalization

> cannot be rephrased in terms of neurophysiology because it explains why the neurophysiology is the way it is. If it did not supply the correct information (to within some approximation), the Darwinian fitness of the animal whose visual system it is would be less.... I might predict that the octopus, or any other sighted invertebrate, would have a VOR, defined as whatever module computes the eye-movement correction factor; I am not predicting anything at all about the neurophysiology of octopi, which is completely different from ours. (p. 180)

The notion of explanation helps to clarify what McDermott is looking for. A question of how the VOR reflex works can be expressed at different levels. While the biological level of explanation may be illuminating, it does not answer the general question of how organisms might keep their eyes oriented toward a fixed target even when their head moves. To answer that question, the VOR reflex needs to be explained in terms of a computation.

McDermott suggests that neurons can be assigned a syntax based on their firing rates. Groups of neurons may be assigned a semantics based on their computational role in a system. It is an objective fact as to whether (to choose an example at random) a cluster of neurons in the retina are computing a difference in light intensities, in the process of detecting an edge. But giving a definite semantics to a computation is fraught with peril. Investigators might surmise that a given computation is involved in the detection of an edge, but this hypothesis can be undermined by what happens "downstream" of the computation in question. If the

result of the computation is never used in other computations involving shapes, as one would expect them to be, then something is amiss with our understanding of the computation.

These are epistemological questions which can be clarified by empirical investigation. A computational theory of consciousness requires that certain computations refer to certain entities or events; determining what certain neural events actually refer to is a different question.

These sorts of explanations, be they of VOR reflexes or edge-detection mechanisms or whatever, are not completely objective. Human interests and values apply constraints to what needs explaining, and what kinds of explanation are acceptable. Our own need to find food and avoid danger makes us interested in how frogs perceive and catch flies, and how they avoid predators. The fact that we also have a visual system means that we will investigate the problem in certain ways, and so on. The semantics of the frog's brain will depend in a hundred different ways on the peculiarities of those who investigate it.

But if it turns out that the question of whether something is conscious has as much (or as little) objectivity as the question of whether a frog's brain helps it catch flies and avoid predators, then I think it is very objective indeed. Since this is all that McDermott seems to require (indeed it may be more than he requires), I would suggest that he has established a sense of semantics that is sufficiently objective to meet Searle's objections.[7]

---

[7] Note that this does not indicate that McDermott and Searle agree. My sense is that for Searle there is a point where any "objective" account of semantics has to be grounded in an individual's subjectivity. Searle would thus complain that McDermott's account is fundamentally incomplete. I am fudging the issue when I say that semantics is "as objective as" science, since similar issues arise there.

In fact, McDermott only needs to make a semantics objective enough that a symbol can be said to denote the system in which it is present. Although there is probably no generally applicable method to do this, it still seems to be more basic and uncontroversial than determining whether a given symbol in a frog's brain means "insect present" or "edible object present." McDermott observes that "just spatial facts about 'I' should be sufficient to single out a unique physical object." (p. 207) I cannot see any reason why this should not be the case.

1.4 The Self-Model

Here's how McDermott explains what a self-model is:

> A computational model C is a computational system that resembles a modeled system S in some respect and is used by a modeling system M to predict the behavior of S. A self-model arises when S=M. (p. 118)

I find McDermott's definition of "computational model" to be a trifle unclear, but I think it may be clarified with an example. Suppose a computer is running a weather simulation. Then the computer is the modeling system M, and the system whose behavior is being predicted is the weather system S. The computational model is a set of numbers (referring to the temperature, humidity and wind velocity of various locations in the weather system), and these numbers are transformed according to certain mathematical rules based on the physics of weather systems. In the ideal case there would be an isomorphism between the computational model and

the actual weather system, but any practical model would be only approximately isomorphic; the model will then merely resemble the thing modeled.

Although parts of the computational model refer to parts of the weather system, this semantic relation is not merely stipulated; McDermott doesn't believe in assigning a semantics, but in discovering the semantics that is "most harmonious." Whatever the intention of the programmers might have been, the true semantics of the model can only be determined after the fact. A highly inaccurate model might not refer to anything.

Saying that a self-model arises when S=M is incorrect, and is in fact contradicted by McDermott's first example (an office furniture inventory that includes the computer doing the inventory among other pieces of furniture). What McDermott should have said is that a self-model arises when $M \leq S$. That is, when the modeling system M is included in what is being modeled.

McDermott's example of the office furniture inventory is interesting because it shows that being a self-model is not a functional notion. That is, something's being a self-model is not defined by its causes and effects.[8] Suppose that one of the employees of that office ran the inventory control program on a personal computer at her home, using the same data. The home computer happens to be physically identical with the office computer, and is loaded with the same programs. It has the same inputs, outputs and inner states as the office computer (and so is functionally

[8] Functionalism claims that mental states are determined by their causes (sensory input or other mental states) and their effects (physical behavior or other mental states). Specifications of functionalism make additional claims, e.g. that mental states are to be identified with computational states (computationalism). My claim is that being a self-model cannot be expressed in causal terms, and so is incompatible with functionalism and its varieties. Or at least for "individualist" forms of functionalism. See the following footnote.

identical) but because the home computer is not in the inventory, there is no self-model. In other words, being a self-model is not determined by a functional description of that individual system

Now since McDermott's thesis is that a mind is a particular kind of self-model, this entails that, for McDermott, the mind is not a functional notion. This seems to contradict McDermott's own wish to defend the variety of functionalism known as computationalism, which he defines as "the doctrine that the mind can be explained entirely in terms of computation." (p. 25) However he notes in the same place that he neglects some of the philosophical issues around functionalism that philosophers like to discuss, so some latitude should be given him as far as definitions go. The position he defends, while only loosely a version of functionalism,[9] is still very interesting.

A quick recap. McDermott's analysis is based on the trichotomy of physical computer, syntax and semantics. Syntax is an abstraction of what is going on at the physical level; unwanted information about the physical states is discarded while something of the structure of the causal relations between the physical states is preserved. Semantics is not stipulated or assigned to the system, but rather the "most harmonious semantics" is discovered by empirical investigation.

McDermott expects that investigation of the brain will reveal groups of syntactic units that refer to particular processes or data and which are used in computations. These groups of syntactic units are computational objects,

[9] My remarks about self-reference and functionalism being incompatible are true, I think, only for narrow, "individualistic" conceptions of functionalism. See Robert Wilson's *Boundaries of the Mind* for an extensive discussion of this topic. I don't think that individualistic formulations of functionalism can account for the limited semantics that McDermott requires to define a self-model.

encapsulated data and processing that provides capabilities to the system. The computation these computational objects are used in is a computational model. The system in this case is modeling its world, and is one of the entities described by the model; the portion of the world model that describes the system itself is the self-model.

Chapter Two

Illusions and Virtual Reality

2.1 Different Kinds of Illusion

If McDermott is right, some systems have minds by virtue of having the right kinds of self-model. But if being a self-model is not a functional concept, a system may be the functional equivalent of a mind but not itself be a mind. We need some terminology to describe the situation when a functional equivalent of a mind is a mind, and when it is not. Jack Copeland suggests (*Artificial Intelligence* p. 47) distinguishing between a simulation$_1$ and a simulation$_2$. A simulation$_1$ is a simulation which lacks something essential to what is being simulated; simulated leather is not leather, because it is made of the wrong sorts of molecules, and a simulated Mona Lisa is not a famous masterpiece (even if it indistinguishable from it) because it doesn't have the right history. A simulation$_2$ is an example of what is being simulated despite the fact that it might be made in a non-standard way; a simulated voice is still a voice, even if it is made by a machine instead of a human, and a simulated protein is still a protein even if it is made in a laboratory instead of in an organism.

Rather than using subscripts, it might be easier to refer to a simulation$_1$ as an imitation, and a simulation$_2$ as a duplication. Alternatively we could say that a simulation$_2$ is 'real' while is simulation$_1$ is 'illusory'. We could thus say of a particular simulation of a mind, itself lacking a self-model, that it is not real, but a mere illusion; an imitation, not a duplication. This is, ironically, the conclusion of Searle, not someone with whom McDermott is inclined to agree. (Although their agreement is mostly coincidental; Searle refers to "causal powers" rather than facts about self-reference when he disagrees with people who believe in strong AI.)

Still, Copeland's distinction doesn't really illuminate the irrealist strand within the work of many functionalist philosophers. Given that our brains are computers which maintain a self-model (and thus real minds), why is it so tempting to describe qualia and free will as illusions? To explain this, a more sophisticated way of understanding simulations is needed.

David Deutsch's *Fabric of Reality* incorporates an understanding of virtual reality that is extremely congenial to McDermott's theory of virtual consciousness. When Deutsch's VR (Virtual Reality) relation is used to express McDermott's notion of a self-model, it comes out that any mind is a VR simulation (in the broad sense that Deutsch uses), and that mental phenomena like free choices and qualia are objective computational objects in the simulation. This independent perspective makes it more understandable why McDermott will often refer to the mind et al as illusions or fictions. Deutsch's explanation of virtual reality also shows that Copeland's division of simulations into imitations and duplications is not exhaustive; while a simulation might be an 'illusion' (a simulation$_1$) or 'real' (a simulation$_2$) there is another kind of

simulation which doesn't quite fit into either category, something I will call a 'real ideal.'

I will explain the virtual reality relation abstractly, and then in the context of a flight simulator. Then I will take some time to draw out this third type of simulation, different from Copeland's simulation$_1$ and simulation$_2$. I suggest that qualia and free will are examples of this third kind of simulation, an illusion that is also real. If I am right this will defuse a lot of the irrealism about mental phenomena that computational models tend to engender.

2.2 The Virtual Reality Relation

Virtual reality in its most general form is a multipart relation. The VR relation has 4 parts: A, B, C and D, and a metric N where

A is the user

B is the simulator,

C is the simulation/illusion/rendered environment,

D is the intended environment, and

N defines how close C is to D.

A and B, in Deutsch's schema, are complex, autonomous realities which interact with each other (i.e. a subset of A's output is B's input, and vice versa). They are usually distinct, but there are important examples where A=B.

C consists in the output of B, which responds sensitively to input from A (usually based on B's stored information and programming) such that the relevant properties of C are close to that of D. For C to be considered a virtual reality, the

inputs and outputs should be unambiguous, and the sense of 'intended environment' needs to be specified. The meanings of terms like "close to" and "intended environment" will have to be given in the definition of N. Note that N has to be (at least implicitly) defined or the choice of D is completely arbitrary.

A and B interact in a sensitive manner, but do not need to process information in the usual sense of the term. For example, an object related to its reflection by means of a mirror is a simple example of virtual reality. The reflected image is only an illusion, an imitation (what Copeland would call a simulation$_1$), a rendering of the thing reflected. The input and output relations are strictly optical, and the closeness relation is one of visual resemblance. Stipulating a broader coverage of the sensory range (or a more substantial element of interaction) would eliminate the mirror from consideration as a VR simulator, of course. Most examples of A and B will involve more complex capacities to process information.

Deutsch does not define the VR relation in this general form. His introductory definition of virtual reality is that "the term refers to any situation in which a person is artificially given the experience of being in a specified environment." (p. 98[10]). He goes on to say that "'virtual reality' is usually reserved for cases where there is both a wide coverage of the user's sensory range, and a substantial element of interaction ('kicking back') between the user and the simulated entities." (p. 99)

An example will clarify the use of these terms. Suppose someone is in a flight simulator. In this case the placeholders are as defined above, with A the user and B the computer which controls the flight simulator. The intended environment D is the cockpit of an aircraft in flight. The 'relevant properties' are the appearance, sound

[10] Unless otherwise noted, all references to Deutsch are found in *The Fabric of Reality*.

and feel of the rendered environment C. The closer C and D are, the less difference there is between how the simulated aircraft responds to the user's actions and how the real aircraft would respond.

It should be obvious that a flight simulator provides only the illusion (a simulation$_1$) of flying the plane, but it might be helpful to examine what makes the experience an illusion. It is not simply that the experience is generated by a computer instead of perceived directly, for a computer generated experience could be real (a simulation$_2$) rather than illusory; the difference between a video camera and a video game. How should we distinguish the two cases?

In the first case, suppose it were necessary to make an airplane cockpit without windows. You might provide a virtual reality window to let the pilot see where the plane is going. This could be a high definition video screen hooked up to some forward mounted cameras in order to give the illusion of a window. If the visual representation of a thundercloud appeared on the screen when and only when a real thundercloud was in front of the airplane, and the real thundercloud was the cause (indirectly) of the visual representation, it seems unreasonable to say that the displayed thundercloud is an illusion. Rather the *window* would be an illusion, but what is seen through the window is real; the simulation is a simulation$_2$.

In the second case, as in a flight simulator, the generated images do not track real meteorological phenomena. Rather, the images are generated according to a program stored in the simulator. The images may be triggered automatically, or by the instructor who is monitoring the simulation, but in either case they do not correspond to reality and so are not real. What is seen on the screen may appear real

and may be mistaken for the real thing, but it is only an illusion; the simulation is a simulation$_1$.

2.3 Real Ideals

It is harder to distinguish between illusion and reality when D is itself unreal. Many putative examples of real ideals can be drawn from geometry; a drawn circle that resembles the ideal, perfect circle and is used in a geometry proof; balls that approximate spheres, dots that approximate points. The idealizations of physics (where all surfaces are frictionless and all vacuums perfect) could also be thought to find their real ideals in the physical situations they explain. But while you could consider the act of drawing a circle to be in some sense "simulating" the (non-existent) perfect circle, it seems more natural to regard the perfect circle as an abstraction and idealization of the drawn circle, and similarly for the other examples. However, even that claim is debatable.

Consider, for example, a line rendered by a computer program. It is straightforward to program the "zoom" function so that the line remains the same thickness no matter the magnification, and come out as perfectly straight no matter what measurement is made of it (as long the tools are those made available by the computer program, of course). Is such a rendered line "real"? I think there could be a legitimate difference of opinion. Someone could say that the image on the screen is physically real, but the image is hardly a geometric line. Among other things, it has a finite thickness. Certain values in the computer's memory may be manipulated by rules corresponding to geometrical transformations, but they are not themselves

geometrical objects. Similar considerations apply for perfect circles, frictionless surfaces, etc. [11] To find an unambiguous example of a simulation which is real though its object is not, we need to work a little harder.

Consider what happens when the VR relation is used to analyze a living organism. Deutsch uses the following justification to support the claim that a living organism is a type of virtual reality.

> I have used the word 'computers' for the mechanisms that execute gene programs inside living cells, but that is slightly loose terminology. Compared with the general-purpose computers that we manufacture artificially, they do more in some respects and less in others. One could not easily program them to do word processing or to factorize large numbers. On the other hand, they exert exquisitely accurate, interactive control over the responses of a complex environment (the organism) to everything that may happen to it... This is more than just computing. It is virtual-reality rendering. (p. 178)

How does this fit the four-part definition of VR? An organism's physical environment fills the role A of the user in a VR environment. The organism itself is the rendered environment C. This is the opposite of what one might expect, for it is the physical environment (not the organism) that is the user and the organism (not the physical environment) that is the rendered environment. So much for A and C. B, the role of the computer, is filled by the cellular mechanism that express the genes, and which (on a phenotypic level) controls the organism. The interaction between A and

[11] cf. *Fabric of Reality* 242-243 for more examples.

B in the VR relation is just the interaction between the physical environment and the genes mediated by the organism.

We have identified A, B and C. What then is the intended environment D, and how is the closeness of the rendering to be determined (i.e. what is N)? This is the key question; without a satisfactory answer the VR analysis of an organism doesn't make sense. Deutsch himself admits that "it makes little sense to speak of a particular situation as being a virtual-reality rendering if there is no concept of the rendering being accurate or inaccurate" (p. 179). His solution is to define accuracy as "the degree of adaptation of the genes to their niche. We can infer the 'intention' of the genes to render an environment that will replicate them." (p. 179).

Deutsch is asserting that genes are VR simulators. It is the "intention" of a VR simulator to generate an accurate simulation. It is the "intention" of genes to be replicated, which occurs when the organisms they code is fit. Deutsch thus equates fitness with accuracy. The most accurate simulation will be the one that is most fit; C is closest to D when C is as fit as possible.

D is the organism perfectly adapted to a particular niche, but while its simulation (the organism C) is real, D is not (in fact, D is probably not even possible[12]). It's the opposite of our usual way of describing an illusion since most often C will be the unreal simulation of an existing D. This situation, where C is real and D is not, is what I call a real ideal. The notion of the real ideal does not seem to fit into Copeland's framework. Copeland assumes that whatever is being simulated is itself real, but now we see that is not always the case.

---

[12] A "perfectly adapted organism" seems reminiscent of a "largest possible number." I don't think it is coherent to imagine such a thing existing even as a platonic ideal.

2.4 Functionalism and Antifunctionalism

By functionalism I mean a theory about the metaphysical status of mental states which denies that mental states have any intrinsic properties. Functionalism holds that the only properties of mental states are causal ones. In this view, an individual's mental states are determined entirely by their causes and effects; some are caused by that individual's sensory stimulation (input), others cause the individual's physical behavior (output) and most are related in complex causal ways to other mental states[13]. By antifunctionalism I mean the position that mental states have additional qualities which are intrinsic and subjective. These qualities make it be like something to be that conscious subject with those states.

Functionalists have two options when faced with the claim that mental states have intrinsic, subjective qualities above and beyond their functional properties. First, they can deny the truth of the statement, typically by stating that features have been erroneously imputed to mental states. Or they can accept intrinsic, subjective qualities as part of the definition of mental states, and then deny that such things exist. "There are no such things as mental states," says this kind of functionalist, "there only appear to be."

McDermott sometimes seems to take the second option; he tends to present mental phenomena as if they were illusory. His discussion is peppered with

[13] This is basic "black box" functionalism. Various suggestions as to how a functionalist theory of the mind might work, and how mental states are to be taxonomized, have been elaborated. But the details of the version endorsed by McDermott (Turing machine functionalism, aka computationalism, aka strong AI) do not protect it against the objection that functionalism can't account for qualia. Nor does the fact that a self-model is computationally equivalent to a Turing machine play an important role in McDermott's explanation of consciousness. I therefore subsume these details into a single story about functionalism.

statements like "a self-model does not have to be true to be useful" (p. 4) and "qualia are useful fictions with a grain of truth." (p. 157). This irrealist strain is not peculiar to McDermott; others who are interested in cognitive science (the highly influential Daniel Dennett is the most famous example) display it as well. Irrealism about the mental is related to the tendency by cognitive science types to say that many (intuitively non-conscious) systems are in fact conscious. It is as if they say "consciousness is an illusion, so why not attribute it to setups like (say) the Chinese Brain[14] that give the appearance of consciousness?"

The first option is hardly better. The features that functionalists claim are falsely imputed to mental states are often what other philosophers regard as essential. Although the claim that mental terms fundamentally misdescribe reality (and so their referents do not exist) is most often associated with behaviorists and eliminative materialists, functionalists like McDermott seem to be tarred with the same brush. When confronted with thought experiments like the China Brain, they are likely to say that yes, the China Brain is conscious. Then, rather than admit that the population of China could collectively implement a mind, some philosophers reject the possibility of any computationalist or functionalist account of consciousness. Instead of saying consciousness is an illusion, they say that machine consciousness is. Searle is the most prominent example of an 'antifunctionalist' philosopher who holds that a computer could achieve only the appearance, but not the reality, of consciousness.

The VR relation makes it straightforward to describe the positions of both the functionalist, "strong AI" camp and their antifunctionalist opponents. More exactly,

---

[14] The "China Brain" is a thought experiment wherein the population of China performs a functional emulation of a human brain. It is intended as a *reductio* for functionalism. (Ned Block, "Troubles with Functionalism")

it describes positions held, on the one side, by various kinds of eliminative materialists and functionalists, who believe in the possibility of artificial intelligence and who tend to be skeptical about the reality of mental states, and, on the other side, the positions held by those who tend to be skeptical of the possibility of artificial intelligence but who believe in the reality of mental states.

This is how to express the two positions using the VR relation. Let B be a computational system, and let A be the same computational system (sc A=B). Let D be a conscious subject with mental states. D has the functional qualities and as many intrinsic or subjective qualities as one's theory of consciousness might require. (The theory might require no non-functional qualities, of course.) Then C is a self-model whose metric N is defined functionally; it is a simulation of D insofar as it is a close functional approximation of D, and it is a self-model insofar as it is also a functional approximation of B.

This relation expresses the antifunctionalist position when B is a computer. According to this position, a computer provides a functional simulation of "real consciousness," but the intrinsic, subjective qualities of consciousness are lost. C is a hollow shell when compared to D. The antifunctionalist states that when B is a human brain it can generate consciousness; C=D and no intrinsic properties are lost. It is important to note that for antifunctionalists D is sometimes instantiated; we can all agree that we have real consciousness.

Functionalists agree with the basic structure of the VR relation, but differ with antifunctionalists about whether D exists. A functionalist says that D is an impossible idealization; only functional approximations to it exist. Consciousness, a

functionalist would say, is an illusion. If the approximation of C to D is close enough, A treats C as if it were D; that's how illusions work. In a good flight simulator the pilot will react to a virtual thundercloud as if it were real; likewise a system will mistake itself as a system with real consciousness.

Although McDermott self-identifies as a functionalist, the position I am presenting as McDermott's is something else entirely. In other words, I argue that McDermott is mistaken when he says he is a functionalist. The notion of self-reference prevents McDermott's theory from being a kind of functionalism, but since his theory doesn't deny the possibility of AI in the way that philosophers like Searle and Block do, it is not an example of what I call antifunctionalism either. To see what McDermott's position actually is, we will first have to express McDermott's notion of the self-model in terms of the VR relation.

2.5 The Self-Model in terms of VR

Consider the special case of virtual reality where an information processing system (ostensibly a robot, but perhaps a human) simulates reality for itself. In the terms of the VR relation, A=B and D is the world. Since the system is a part of reality, its model of reality contains a model of itself.

Recall McDermott's definition of the self-model:

> A computational model C is a computational system that resembles a modeled system S in some respect and is used by a modeling system M to predict the behavior of S. A self-model arises when S=M. (p. 118)

I understand McDermott's definition to involve three entities, C, S and M, where C resembles S; C is used by M to predict the behavior of S; and M is a subset of S[15]. This is a variation of the virtual reality relation explained previously. M uses C, so M is the user. C is used to predict the behavior of S, which it resembles, so C is the simulation and S is the intended environment. M is a subset of S, so the intended environment includes the user. Since M is described as the modeling system, it must generate the simulation; thus it is also the simulator. In other words, a self-model arises when A=B, and the simulation includes a model of the entity who is performing the simulation. Since the systems we will be discussing are capable of sense perception and physical actions, D is at least a subset of the physical world, the arena where sense perception and physical action take place.

McDermott's thesis is that the mind is a certain kind of self-model, one that (among other things) models the organism's perceptual events. McDermott's discussion of the self-modeling of perceptual events leads to an account of qualia that says they are features of the simulation C, not of the world D in which the organism finds itself.

To flesh out the implications of McDermott's theory of virtual consciousness, it is helpful to incorporate Deutsch's virtual reality theory of perception. Deutsch's take on sensory perception reinforces and supplements McDermott's virtual consciousness theory. Deutsch describes perception as a virtual reality computation implemented in the brain. This description reiterates very forcefully the fact that the virtual reality simulation must be generated by the entity whose perception it is. This is another way in which the semantics of the VR relation (and not merely their

[15] Recall that I argue that M≤S should be read in place of M=S.

functional role) constrain what counts as a mind. Not only must a mind consist of a self-model, it must consist of a self-generated VR simulation.

2.5.1 Sense Perception as VR

Deutsch argues that our phenomenal experience of the world is a type of virtual reality. He says

> ...our external experience is never direct; nor do we even experience the signals in our nerves directly – we would not know what to make of the streams of electrical crackles that they carry. What we experience directly is a virtual-reality rendering, conveniently generated for us by our unconscious minds from sensory data plus complex inborn and acquired theories (i.e. programs) about how to interpret them. (p. 120)

Deutsch seems to be using a different schema here than was described earlier, but the difference is merely notational. He is restating the VR relation as "B provides A with a simulation C of an external reality D." B is the 'unconscious mind', which uses 'nerve-crackles' from the sense organs (which in turn are based on events in the outside world D) to provide an experience of D. A is both the subject of conscious experience and the source of action which has an effect on the world. It is a conscious subject insofar as its input is B's output. It has an effect on the world because its output (nerve-crackles leading to muscles which result in movement and action) changes the world in a way that produces an effect on B. The closeness relation N (between C, the perceived world, and D, the real world) seems to be

pragmatic; the world must resemble our perception of it closely enough that we can negotiate it with confidence. However we know that C and D do not match completely. Uncontroversial examples abound; optical illusions, for example.

According to this model, visual perceptions are generally real and veridical. They track the real state of affairs in the world, and are caused by the state of affairs they correspond to. There are exceptions, since we are subject to optical illusions and hallucinations, but these occur precisely when our visual perceptions no longer track the world, or are no longer caused by the things perceived. If you perceive a dog in front of you when there is no dog, you are experiencing some kind of hallucination. If by coincidence there happened to be a dog in front of you at that very moment, but it was not the cause of your perception, it still counts as a hallucination.

2.5.2 Internal Experiences

Although Deutsch applies the VR model to sense perception- what he calls "external experience"- he hesitates to discuss VR simulations of "internal experiences" like emotional responses and mood. Deutsch says (pp. 103-104) that a machine which can supply these experiences would override the user's mind as well as the user's senses; it would thus be replacing the user by a different person. He considers the case where only sensory experiences are involved, where A and B can be distinct without B affecting A's personhood. If emotional experiences were involved, B would affect A's personhood.

This suggests an answer to Robert Nozick's famous question about the experience machine. Nozick asks us to

> ...suppose there were an experience machine that would give you any experience you desired. Superduper neuropsychologists could stimulate your brain so that you would think and feel you were writing a great novel, or making a friend, or reading an interesting book. All the time you would be floating in a tank, with electrodes attached to your brain. Should you plug into this machine for life, preprogramming your life's experiences? (*Anarchy, State, and Utopia*, p. 42)

Deutsch's response would be that using the experience machine would be a kind of suicide; your mind and personality would be replaced by a construct of the machine's, and you would be lost, perhaps never to return. Nozick's follow-up question about "a transformation machine which transforms us into whatever sort of person we'd like to be" (*Anarchy, State, and Utopia*, p. 44) would be rejected for the same reason. Nozick supposes that our using such machines would be "compatible with us staying us" (p. 44) but Deutsch would disagree.

McDermott, on the other hand, uses a model that is equivalent to the VR model to explain experiences of qualia, emotion, and free will. These are internal experiences, and so threaten the integrity of the user's mind if they are supplied from the outside. In the cases I consider, however, these experiences are self-generated by the user; that is, A=B and so Deutsch's reservations do not apply.

This ready identification of A and B might seem too glib. Doesn't Deutsch suggest that a distinction can be maintained when he distinguishes the unconscious mind, which generates a VR rendering from sensory data, and the conscious mind,

which actually experiences the VR rendering? In other words, might not A and B inhabit distinct parts of the brain? Unfortunately, making such a distinction between A and B leads to difficulties. Let's see why.

2.5.3 Why A and B shouldn't be distinguished

First, let's back up and reconsider the example of the flight simulator. Here it is easy to distinguish A (the user) from B (the simulating computer), since a bulky interface of headphones, video screens and motion sensors separate them. But what if this interface was bypassed, and B interacted with A's nervous system directly? (See *Fabric of Reality* pp. 108-109) B takes as input the nerve signals A sends to A's muscles. B generates output which directly stimulates the nerves in A's inner ear, retina, and so on. A certain amount of processing power is required to calculate the movements that the motor neurons would produce, and to calculate how the neurons would respond to sensory input, but there seems to be no physical reason to assume this couldn't be done. But what practical reason is there to stop here?

The human brain contains roughly 10^{11} neurons. Many of them connect directly to sensory organs, muscles or glands. These are "first-layer" neurons, and are what B connects to, having bypassed the muscles and sensory organs. The neurons that the first-level neurons connect to are "second-layer" neurons, which connect to third-layer neurons and so on (this is a structural description; physically the layers may be interwoven). Since several thousand neurons are in the first layer and each neuron connects with approximately 100 to 1000 other neurons, a total of less than ten layers should be sufficient to account for all the neurons in the brain; maybe only

four or five. But for the purpose of the argument, the exact number is irrelevant; there could be a hundred layers or more.

What happens if, in addition to bypassing the muscles and sense organs, B bypasses the first layer of neurons as well? The functions of these neurons will have to be simulated, of course, but if this is done correctly the machine can interface directly with the neurons of the second layer. But if the first layer of neurons can be bypassed, why not the second, or the third? Soon all the neurons have been bypassed, and nothing of A remains. Without A there is no interaction with B, and without an interaction there is no simulation C. In Deutsch's model the mind seems like the core of an onion; you can search for it by peeling away the layers, but in the end nothing remains.

Deutsch allows first level neurons to be replaced, but prohibits the replacement of neurons involved in "inner experiences" (like emotional responses) that presumably are handled at a deeper level. This prohibition is one way of stopping the "onion objection" before it gets very far, but won't work for Deutsch's account of how the 'unconscious mind' generates a VR rendering of the world on the basis of the 'nerve crackles' which the sensory organs convey to the brain. Here it seems that the first layer of neurons are outside of the brain; it is their nerve-crackles which is the input for the rendering. What Deutsch calls the unconscious mind must consist, at the very minimum, of the second layer of neurons. But this second layer itself produces nothing but nerve crackles, which on Deutsch's assumption is incomprehensible. By repeating the argument we conclude that it is "nerve-crackles all the way in," and we find nothing at the core of the onion.

The only solution to this puzzle seems to be to stop it at the first step. I suggest that in the VR model of the mind it is not helpful to distinguish between A and B. The organism as a whole perceives its world by generating a VR simulation with its sense organs and nervous system, and it consults and interprets this virtual reality as a whole organism. The content of its understanding is nothing but the content of the simulation. There is no fact of the matter where the boundary between A and B lies, and no importance should be attached to the question. Deutsch's distinction between the conscious and unconscious mind, I suggest, comes out of the way that the system models its own perceptual, cognitive, and volitional processes; the features that are modeled are conscious, and the unmodeled features are unconscious.

The essential machinery is in place. We have McDermott's model of virtual consciousness expressed in terms of a virtual reality simulation. Now it is time to discuss qualia.

Chapter Three

Qualia

3.1 What are Qualia?

The Cambridge Dictionary of Philosophy (2nd edition) defines qualia as "those properties of mental states or events, in particular of sensations and perceptual states, which determine 'what it is like' to have them."

There is, in addition, something fundamentally irreducible about qualia. Show a friend a ripe tomato, and ask what it looks like. "Like a tomato." should be the response. If you ask, "What do you mean? How does it look like a tomato?" your friend should explain how it is red and round and bulgy- like a tomato. If you ask how they know it is red and round and bulgy, they should be at a loss for words; no matter how much they think about it, the experience of redness or roundness or bulginess cannot be further reduced (though they may be paraphrased: words like "curved" and "convex" might be substituted for roundness and bulginess). That is what I mean by the irreducible aspects of qualia; those aspects of a sense perception, like redness and roundness and bulginess, that cannot be further reduced.[16]

[16] To be clear I should emphasize that I am not interested in explaining these verbal reports unless the explanation is in terms of what these verbal reports refer to; the qualia of the reporting agent.

To understand qualia, it is important to note that sense data have two aspects; a system can take them as reports of the state of the world, but a more sophisticated system could also take them as reports of the functioning of the system. For example, a particular visual experience can be taken as the perception of something that happens to be red, or it can be taken as the experience of redness that happens to be caused by something.

This distinction is important if sense data are misleading or ambiguous. McDermott's example (p. 106) is of a stick partially submerged in water. Visual reports indicate the stick is bent, but tactile reports (from handling the stick) indicate that it is straight. To decide whether to model the stick as straight or bent, the system needs to distinguish the sense data from the thing sensed; more concisely, the system needs to distinguish appearances from reality. The system's model of the world needs the ability to model qualities independently of the particular objects they are qualities of.

This is an important distinction because it is quite conceivable for an animal or robot to have the ability to discriminate sensation without independently modeling qualities. A system with this more limited capability will be able to respond appropriately to many situations, but will be unable to recognize illusions, negotiate new situations, or uncover deception. These kinds of capacities require one to distinguish between how things appear to be and how they really are. I will assume that any system we are interested in will have the ability to sense its environment and react appropriately to new situations. Such a system will, among other things, be able to model qualities independently of the objects they are qualities of. It is

McDermott's contention that qualia arise in exactly this situation; when a system can distinguish between an object being a certain way and only appearing that way.

A quality that is independently modeled typically has two associated pieces of information. First is its sensory modality; whether it is perceived by sight, like color is, by touch, as warmth is, and so on. The second piece of information compares it to known standards, and assigns it to a place in similarity space. For example, a colored object might be found to be rather like a well-watered lawn and a little bit like a sunny sky, and a tub of water might be described as being somewhere between warm and lukewarm. These qualities might be assigned adjectives; "greenish blue", say, or "tepid". Descriptions of qualities usually have both pieces of information; something might smell like lemons, or sound like a flute.

The simple kinds of qualities that I want to describe are determined by a sensory modality and one or more comparisons; I intend to neglect other complexities of sensory experience. If a system can determine if something looks more like red than orange or smells like garlic, then it has the discriminative ability of a system that has unquestionably real experiences of qualia. A model that can handle simple qualities can probably (with minimal modifications) handle ones that are more richly structured. Examples of a richer structure for the quality of color would be a determination of whether a color is stimulating or relaxing, warm or cool, attractive or repulsive; it might also include emotional resonances determined by particular experiences associated with that color. Systems should also distinguish stimuli that are remembered from ones that are currently being perceived, but I shall neglect that distinction too; I will presume that all stimuli are current. Finally, I will ignore

phenomena like optical illusions. A given sensory system may have quirks, but I am interested in its ordinary functioning, not when it goes wrong.

Systems with linguistic ability will be able to report the sensory modality and the result of a stimulus comparison- that is, they will be able to describe qualia- and in the following discussion I will presume such linguistic ability. This is simply to avoid awkward paraphrasing for systems that can make the appropriate discriminations and modifications of their behavior, but can't actually talk.

Now the fact that a system can report the sensory modality of a perceptual event seems straightforward; the stimuli are coming along particular nerves (or in the case of a robot, along particular wires) and the system can record this fact in its self-model. But what is the basis of a comparison judgment? If the system were asked how it knew that an object was an optical match to an exemplar of green, how could it respond? It would say "it just looks green- see for yourself!" but that doesn't really answer the question.

In humans, color perception is based ultimately on the differential response to different frequencies of light by the cones in one's retina, and this differential response is based in turn on the chemical properties of the visual pigments in the eye. But facts about one's cones or visual pigments are not perceived directly; before the advent of the microscope no-one even knew we had such things in our eyes. All we knew is that some things looked green and other things looked red, and so on. Since the invention of the microscope our theories have improved, but our senses remain the same. It is still the case that our perceptions "bottom out" long before the cellular level is reached.

Whatever mechanism a system uses to perceive qualities, it too must bottom out eventually. McDermott describes the situation as follows:

> Just as people have no introspective access to the fact that their color judgments are based on the differential sensitivity of three visual pigments, the robot has no access to the equivalent fact about itself. (p. 113)

Adding introspective capacities to the robot won't help matters. Suppose that the robot had a human retina, but that in addition the retina was fitted with a micro-miniaturized array of neural probes capable of sampling the activity of each cone, rod and neuron. If asked why something appeared red, the system could refer to the self-model and give a detailed neurophysiological answer; these cones are more active than those, this pigment is responding more to this frequency of light than this other pigment is, these are neurons are being stimulated while those are being inhibited, and so on. But if the system were asked how it knows that certain cones are more active than others, how could it respond?

Its response would depend on how these facts are represented in the self-model. The system might be able to say that those cones feel/look/sound more active, but the basis of these perceptions would be beyond it; the answer to the questions could not be given based on what is in the self-model. (Unless it had yet another system with which to monitor the activity of the neural probes; but this leads to an infinite regress.) It is quite possible that it would not be able to report even that much; it might "just know" something. After all, how do you know that a color is perceived visually? You just know.

So given that the information in the self-model about the perceptual system must bottom out, what can a system report about the most basic kinds of perceptions? It could say what perceptual sub-system they are reported in; this is the sensory modality. And it could make comparisons; that is, it could say whether two samples match. If two samples match visually, it would report that they "look the same." If one sample was an exemplar of green, it could simply say that other sample "looks green." Similarly, if two stimuli are matched in the auditory system, they "sound the same." If they are olfactory stimuli that don't match, they "smell different" and so on. Individual stimuli might be the subject of any number of comparisons, but it is clear that the system can report nothing else about these 'bottomed-out' sense perceptions other than their modalities and the results of the comparison(s) with other stimuli. These bottomed-out features of the self-model that enable these reports to be made are, I suggest, nothing less than qualia themselves.

This explanation of qualia accounts for the main properties identified above. Qualia permit only comparisons, not identifications; they allow a system to say what something is like, but not what it is; this is the qualitative nature referred to above. They are irreducible because the neurological events that are used to generate the qualia are not accessible to the system.

3.2 Chalmers and McDermott

In *The Conscious Mind* David Chalmers anticipates McDermott's treatment of qualia. In fact, McDermott's book is a response to David Chalmers' book, which McDermott calls a "great, if totally misguided, book on consciousness that seemed to

demand a response" (p. xv). In *The Conscious Mind* Chalmers anticipates the basic insights of McDermott's theory (see pages 184 to 189 and 288 to 292) and says "if I were a reductionist, I would be this sort of reductionist" (*The Conscious Mind*, p. 189). However Chalmers regards McDermott's kind of reductionism as being a kind of eliminativism, and sees this as untenable. No matter how convincing the argument that there is really no such thing as consciousness or qualia (there only appears to be), Chalmers maintains that we still have conscious, qualitative experiences.

The following excerpt illustrates Chalmers' anticipation of McDermott (all italics are due to Chalmers):

> ...imagine that we have created computational intelligence in the form of an autonomous agent that perceives its environment and has the capacity to reflect rationally on what it perceives... such a system would surely have some concept of self... [it] would be able to access its own cognitive contents much more directly than it could those of others... [and] would most naturally have direct access to perceptual information, much as our own cognitive system does.
>
> When we asked the system what perception was like, what would it say? ...It seems... likely that it would say, "I know there is a red tricycle because I *see* it there."... When we ask how it knows that it sees the red tricycle, an efficiently designed system would say, "I just *see* it!" When we ask how it knows that the tricycle is red, it would say the same sort of thing that we do: "It just looks red." If such a system were reflective, it might start wondering about how it is

> that things look red, and about why it is that red *just is* a particular way, and blue another. From the system's point of view it is just a brute fact... Of course from our vantage point we know that this is just because red throws the system into one state, and blue throws it into another; but from the machine's point of view this does not help. " (*The Conscious Mind*, p. 185)

Chalmers does not think that an explanation of this sort is sufficient to explain phenomenal consciousness. He admits that such an explanation could explain "why we say that we are conscious, and why we judge that we are conscious" (*The Conscious Mind*, p. 186) but Chalmers thinks that consciousness itself is not being explained. For "it is at least *logically* possible that one could explain the judgments [about consciousness] without explaining consciousness" (pp. 188-189) and so consciousness is conceptually different from what is explained by this kind of explanation. Consciousness, says Chalmers, "is a brute explanandum, a phenomenon in its own right that is in need of explanation." (p. 188)

Concerning this point McDermott writes, "the flaw in this line of reasoning should be obvious. All our theory needs to do is explain why consciousness seems like a brute explanandum." (p. 147)

Now *Mind and Mechanism* was published in 2001, five years after the publication of Chalmers' *The Conscious Mind.* It would seem that McDermott should be able to respond to Chalmers' book but not vice versa. But this is not the case; Chalmers rejects McDermott's point in the following passage:

> We might even reductively explain why I think conscious experience is an explanandum. This might be thought to undercut my arguments in earlier sections entirely, opening the way for a reductive view of consciousness. But again this view can be satisfying only as a kind of intellectual cut and thrust.... An explanation of behavior or of some causal role is simply explaining the wrong thing.... The puzzle of consciousness cannot be removed by such simple means. (*The Conscious Mind*, p. 189)

There are two parts to Chalmers' reluctance to accept the sort of explanation he sketches and which McDermott expands on. The first is that such reductionism seems to entail an eliminativism about the mental which is totally alien to our everyday experience. It is as if a philosopher "proved" that we didn't have hands. It is partly to counter this aversion to eliminativism that I have emphasized the fact that just because something is an illusion (in the sense of being generated by some kind of VR simulation) doesn't mean that it is not real. McDermott's explanation of qualia doesn't mean we should stop believing in them; they are perfectly real computational objects whose natural environment is the VR simulation of the world that brains generate for us.

The second part of Chalmers' reluctance lies in his understanding of what is being explained. He argues that since it is logically possible to explain the judgments we make about consciousness without explaining consciousness, then consciousness is conceptually distinct from the judgments we make about it. When Chalmers says consciousness he means phenomenal consciousness, and so he is talking about qualia.

I would argue that McDermott's theory of virtual consciousness provides a theory of what qualia are, not just a theory of how to explain the judgments we make about perceptual events, and that it therefore is in fact a theory of consciousness.

What Chalmers did not anticipate was McDermott's innovative idea that the self-model is constitutive of mind. Chalmers himself speculates that information is the fundamental "mind-stuff," a notion that leads him to a kind of pan-psychism. McDermott's theory is not nearly as profligate in its attribution of consciousness.

3.3 Other Perspectives on Qualia

Daniel Dennett's remarks about visual perception are quite compatible with the virtual consciousness theory:

> What science has actually shown us is just that the light-reflecting properties of objects- their secondary qualities- cause creatures to go into various discriminative states, underlying a host of innate dispositions and learned habits of varying complexity. (*Brainchildren*, p. 143)

The discriminative state referred to is the state of the self-model; how the system is modeling its own perceptions. These states underlie the reports the system makes and the behavior it subsequently engages in; they are the basis for qualia.

E. Conee identifies the experience of qualia with an acquaintance relation:

> Learning what red things look like is identical to learning how red things look, and this is identical to learning the look of red things... Clearly this does not say that what is learned is some fact to the effect

> that something or other is so. It says, concerning a certain look, that what is learned is *it*. A look is not a fact. This learning seems to be unproblematically classified as a relation of the person to a phenomenal quality, just as the acquaintance approach would have it. ("Phenomenal Knowledge", p. 142)

Conee speculates that "attentively experiencing a quality might be a brain state with a complex neurophysiological nature." ("Phenomenal Knowledge", p. 147) In this he echoes P. M. Churchland who says that acquaintance knowledge may involve a special "prelinguistic or sublinguistic medium of representation for sensory variables." ("Reduction, Qualia, and the Direct Introspection of Brain States," p. 23) Qualia are modeled in the special medium that Churchland refers to, and their non-propositional representation there is a way of satisfying acquaintance relation that Conee specifies.

Earlier in *The Conscious Mind* Chalmers notes that qualia are subjective, that is, defined in terms of a subject (p. 4), that they are qualitative, "like something" (p. 4), and that they are always accompanied by awareness, that is, by the "state where we have access to some information, and can use that information in the control of behavior." (p. 28)

Because qualia occur only when an entity has a sophisticated self-model and a self-generated simulation of its world, they depend on that entity for their existence; this accounts for the subjective nature of qualia. The virtual consciousness theory explains the qualitative aspect of qualia: since the base neurochemical events are inaccessible to the system, only comparisons can be made. The virtual simulation of

the world is generated so that the system can predict and respond appropriately to the world, so it makes sense that the features of the simulation are always associated with awareness; the access to information that is needed for the control of behavior.

3.4 Contentious Qualia

There is another kind of qualia which I will call contentious qualia. These are qualia which can be inverted or removed without there being an associated functional change in the system. I get the name from the following excerpt of Dennett:

> The contentious sense [of qualia] anchors the term to the presumed possibility of 'inverted qualia' and 'absent qualia.'... In that sense, there are no qualia. There is no other sense that has a clear and agreed-upon meaning, so I have recommended abandoning the word, but I seem to have lost that battle. (*Brainchildren*, p. 141)

The notion of inverted qualia is based on the logical possibility that someone else might experience a world that is qualitatively very different from yours. Where you see red, they see green, and vice versa; and this is logically possible (or so it is claimed) even if you are physically identical.[17] Such people would have learned that grass is referred to with the word "green" and that roses are called "red." But when they say these words they refer to the opposite internal sensations.

---

[17] One might legitimately wonder if other people have relevant physical differences that lead them to perceive the world in quite different ways. Suppose someone had their photo-pigments reversed: the green photo-pigment in their R-cones (the cause of one kind of red-green color blindness) and the red photo-pigment in their G-cones (the cause of another kind of red-green color blindness). Such a person would have "pseudonormal" vision. It certainly seems plausible that they experience an inverted spectrum. This case raises various difficulties which cannot be dealt with here. See Martine Nida-Rümelin's "Pseudonormal Vision: An Actual Case of Qualia Inversion?" for more details.

The notion of absent qualia is reminiscent of the problem of other minds; it is based on the logical possibility that someone else might have no qualia at all; it is like nothing at all to be that person, despite the fact that they may be physically identical to you and behave exactly like you do. People with absent qualia are called "zombies."

Recall the example of two identical computers running the same inventory program. One computer is included in the inventory, and so has a self-model; the other isn't and doesn't. If two systems were physically (and functionally) identical, could not the same thing happen? One system would have a self-model and a mind, but the other wouldn't. This would appear to be an example of zombiehood. The problem with this line of reasoning is that a computer's presence in an inventory is a very feeble kind of self-modeling. A system like a human person is such that a physical duplication would result in the creation of another person; as McDermott suggests, the physical facts alone would ensure self-reference.[18] What happens if we assume that any interesting system is such that any physical duplication has a self-model?

In this case, according to the virtual consciousness theory, if one such system has qualia, a physically identical system (which it must be, if it is physically identical) must have qualia too. If the virtual consciousness theory of perception is correct, sense experience is the generation of computational objects in the system's self-generated virtual reality simulation of the world; and these computational objects are qualia. These computational objects also play a role in the guidance of behavior

[18] A similar assurance would not seem to be present in the case of Searle's Chinese Room or Block's China Brain. Perhaps these are cases where functional duplication results in only the imitation of consciousness.

(such as the generation of verbal reports about what is sensed), but that is a different aspect; awareness, not consciousness. But if they were absent, the system would lack the ability to make verbal reports about sensory experiences. Since by hypothesis the two systems behave the same, then these computational objects are present, and so both systems have qualia.

Inverted qualia are equally impossible in the model, at least if the two systems are physically identical. Although first there is the question as to whether qualia can, even theoretically, be compared between subjects. If A's qualia summarize comparisons of A's perceptual events, and B's qualia summarize comparisons of B's perceptual events, on what basis should they be compared? They are about different things, after all. If they are compared on the basis of their syntactic representation or functional role then, given that A and B are physically identical, the qualia must also be identical.

I would argue, then, that qualia in this theory are not "contentious." However, note that the reference of qualia are determined by the scientifically ascertained "most harmonious semantics." Perhaps there is a scenario where the best explanation assigns different semantic values to the qualia in physically identical systems; red to one, and green (or even nothing at all) to the other. I myself have not been able to formulate such a scenario.

Perhaps, one might wonder, the qualia described in the virtual consciousness theory are not "real qualia," but are only clever imitations? This seems to be a difficult claim to make. If there were some reason to think that the presence of real qualia were due to some special power or innate capacity, then this special power or

capacity could plausibly be held to be unique to the human race. The gift of some benevolent deity, perhaps. A gift that could not be presumed to be present in any human construct, no matter how clever it appeared to be. You don't assume they have a particular capacity without proof; rather, you assume they are incapable unless shown otherwise. But that is exactly the case with qualia; they are due to the ineliminable incapacity for any finite being to model itself completely, and so the presumption should be that an otherwise talented robot possesses this incapacity.

The qualia of the virtual consciousness theory are real. For them, being modeled makes them real. Although these computational entities do not exist in the physical world outside the simulation, they are perfectly real syntactic units in that system's computational model of the world, and they have a perfectly objective semantic relation to the sensory events they are associated with.

Chapter Four

Free Will

4.1 McDermott's Theory

McDermott's treatment of free will is to some degree parallel with his treatment of qualia. Both phenomena are explained by reference to how a system models itself, and both can be explained in terms of virtual reality. Whereas qualia result from the way that perceptual events are modeled, free will results from the way that decision making events are modeled. The models of perceptual events and of decision making both "bottom out" with parallel results. Just as his discussion of qualia focuses on the subjective and qualitative aspects of perception, so McDermott's discussion of free will is concerned with the subjective and qualitative aspects of decision making. His theory of qualia is about the experience of perception; his theory of free will is about the experience of decision making.

McDermott's theory of free will is tricky to classify in conventional terms. Most theories try to reconcile the existence of free will with the fact that physical processes are determined (at least probabilistically). Since McDermott's theory is that free will is an illusion, it would seem natural to classify him as a hard determinist. But the most natural reading of his theory of qualia would lead one to

classify him as an eliminativist; a reading I argue is mistaken. I argue that it is similarly mistaken to classify McDermott as a hard determinist. In fact, it may be mistaken to classify him as having any particular theory at all about free will.

McDermott points to a contradiction between what the physical nature of the universe entails and what one experiences. McDermott explains that physical processes unfold according to causal laws, and that a "causal law enables one to infer, from the state of a system in one region of space-time, the states at other regions, or at least a probability distribution over those states." (p. 7) However, this does not correspond to what people experience when they make choices. They feel that the future is open, a mix of possibilities that are up to them, not a single possibility (as determinism would have it) nor a probability distribution (as indeterminism would have it). Since this experience is at odds with the physical nature of the universe, the experience of free will is therefore an illusion. An illusion of what? McDermott says that free will is the illusion of being exempt from causal laws. It is his contention that this illusion arises when a system can change a decision based on the consequences it foresees for that decision.

Now there are many different takes on the problem of free will, but hardly anyone thinks that being free means being exempt from causal laws[19]. If our experience of free will means feeling exempt from causal laws, then this is an illusion to be accounted for no matter whether you are a hard determinist, compatibilist or

[19] Libertarians (folks who claim that an event's causal antecedents do no always uniquely determine what happens) might claim to be exempt from causal laws, but most don't. The physical story behind a libertarian position typically involves quantum indeterminacy and chaotic systems, and these are subsumed under "causal laws" as McDermott defines them.

libertarian. McDermott's theory thus supplements any of these other theories of free will.

What exactly does he argue? McDermott argues that free will lies not in the fact that some beings are exempt from causal laws, but rather that some beings cannot help but think that they are exempt from causal laws. He says that

> Any system that models its own behavior, uses the output of the model to select among actions, and has beliefs about its own decisions, will believe that its decisions are undetermined. What I would like to claim is that this is what free will comes down to:
>
> > A system has free will if and only if it makes decisions based on causal models in which the symbols denoting itself are marked as exempt from causality. (p. 98)

I interpret this excerpt as one complex claim, not two; the second paragraph summarizes and gives a term for the situation described in the first paragraph. I also understand that something is "exempt from causality" and "undetermined" if and only if its decisions are not subject either to a unique or to a probabilistic determination. When this is the case, the system must treat the decision making process as a black box. These two paragraphs can thus be restated as follows:

> If a system
>
> 1. Models its own behavior in terms of causes
> 2. Uses the output of the model to select among actions, and
> 3. Has beliefs about its own decisions
>
> Then,

X. It will believe that its own decisions are undetermined,

Y. It will mark the symbols denoting itself as exempt from causality, and

Z. This is all that is meant by free will.

The premises that indicate the kind of self-model the system must have are hard to satisfy; McDermott is not interested in systems with simplistic models. The system's model of the world is sensitive to the system's own causal role in that world and incorporates causal knowledge (supporting the relevant counterfactuals) in sufficient detail to model the outcome of its actions. In other words, the conclusions do not result trivially from an easily repaired flaw of a false or inadequate causal model. Another way to make the same point is to say that all possible models that satisfy the premises are inadequate in the sense that they admit the illusion of free will. In any case, McDermott is concerned with the best causal models that include the system doing the modeling.

Note that McDermott is not simply describing the situation where a system, failing to track some of the causes of its behavior, models its behavior as coming from nowhere. This would be an error in modeling which could be repaired by supplying more information and/or computational capacity to the system. McDermott's argument is that any system, no matter how well-informed, will have the illusion of free will if it can change a decision based on the consequences it foresees for that decision.

Of the conclusions, Y is a practical consequence of X; if the system's decisions are undetermined even in the best causal models, the system needs to keep track of this fact. Z is a definition of free will "from the inside"; it states that the

experience of freedom lies in one's decisions being undetermined according to any useful causal model of one's own behavior.

Here's what I mean by useful. A deterministic physical model is not useful if it requires a system to model itself in complete detail; as McDermott shows (p. 97), doing so would involve an infinite regress. A 'model' consisting of the statement "all things have causes" is not useful either; it cannot be used to select among actions. However, even with these considerations McDermott's argument needs to be supplemented before his conclusion follows.

4.2 The Robot and the Bomb

To defend his claim, McDermott tells a story of a robot sufficiently sophisticated to model the world in causal terms. McDermott's requirements are somewhat modest; first among them is that the robot's model of the world includes a symbol R which denotes itself. By self-denotation McDermott means that

> ...when it detects an object in its environment, it notes that R knows the object is present; and when it has a tentative course of action on hand, that is, a series of orders to be transmitted to its effector motors, it will base its modeling activity on the assumption that R will be carrying out those actions. (p. 96-97)

This means that the robot interacts with the world via a VR simulation, and the VR simulation includes a self-model. The VR simulation is one where A=B. No special perceptual abilities are required, but the robot is able to sense its surroundings.

Now the robot happens to be in a room that includes a bomb with a lit fuse, and it has no plans to leave. Its model of the world includes the fact that R is in a room that includes a bomb with a lit fuse, and that R's action list includes no plans to leave. Based on these factors, the robot will predict that R will be destroyed.

However, this prediction triggers a standing order in its programming to avoid damage. The standing order states that, in the event that its self-model predicts the destruction of R, the robot is to discard its current action list in favor of actions which will protect it. The robot does so, and decides to exit the room.

All this is obvious to an external observer. As McDermott states, "The sequence I laid out is a straightforward causal chain, from perception, to tentative prediction, to action revision" (p. 97) But McDermott argues that it is far from straight-forward from the robot's own point of view. Its prediction of the future is based on a model of its own future behavior. However its future behavior depends on what it predicts in the present. The robot seems to be in a catch-22 situation.

4.3 Supplementing McDermott's Argument

McDermott's account is incomplete. One element that he leaves out is a kind of variance condition for the available actions. Among the courses of actions available to the robot, some must lead to different outcomes. Without

a variance condition we have no way of representing how the robot could both be unfree and know that it is unfree.

Imagine, for example, that the robot was thrown out of an airplane. The robot analyzes the situation and determines that if R does nothing, it will be destroyed on impact. The standing order to avoid damage is activated, and the robot begins to consider different actions. Unfortunately the model of R does not include any built-in parachutes or braking rockets, nor is there any radio transmitter with which to summon mid-air rescue vehicles. In fact, no action of R will change the outcome of the model; R is doomed. Under these circumstances it seems meaningless to conclude that the robot has free will; it is not free to not fall. The robot cannot believe it is exempt from causality.

So let's read McDermott generously, and understand condition 2 as referring to actions that make a difference. The robot, while plummeting, has no actions available which will change the output of the model. Condition 2 is not satisfied, so the conclusions do not hold, or hold only vacuously.[20]

Even if you read meaningful actions into condition 2, the strongest conclusion that seems to follow from McDermott's story is that the model of R's behavior is always revisable. How does McDermott argue his (much stronger) conclusion?

McDermott focuses in on a peculiar feature of the robot's decision making procedure; namely that "some of the causal antecedents of R's behavior *are situated in the very causal-analysis box* that is trying to analyze

[20] After all, if you can make no decisions, it is vacuously true that your decisions are undetermined. But the incapacity to make meaningful decisions seems to be a strange thing to equate with the possession of free will. So it seems better to say that the conclusion does not hold.

them." (p. 97, emphasis McDermott's). He argues that for this reason the causal chain that leads to the robot's departure from the room cannot be accurately represented by the model, because "it would have to include a complete model of itself, which is incoherent." (p. 97) It would require infinite resources to model itself modeling itself modeling itself modeling itself....

4.4 Refuting McDermott's Argument

McDermott's claim is that an accurate self-model leads to incoherence when the causal antecedents of the robot's behavior are situated in the causal-analysis box that is trying to analyze them. This is true if to be accurate the model had to be complete; that would involve an infinite regress. But is it true that accurate representation requires a complete model? Early on McDermott establishes the principle that "a self-model does not have to be true to be useful." (p. 4) This suggests that the robot might not need a completely accurate model of itself; a model that is arbitrarily close should be close enough. It is not hard to show how a robot in the bomb situation could model itself with a sequence of ever-closer approximations, each of which yields the same prediction of the robot's future action. But if this is so, then R's departure from the room can, in fact, be modeled accurately by the robot.

Let us see how this self-modeling could work. In order to minimize confusion in what follows, it will be helpful to have a name for the robot which is distinct from the symbol R it uses for itself. Call it Robbie.

Suppose Robbie models a robot R_0 such that R_0 perceives that a bomb is present in its environment, tentatively predicts its own demise, and exits the room, all as described above. This is a causal-epistemic model; a causal model that includes an account of what one or more agents know. The model contains information about the intensity of the explosion, R_0's structural integrity, that R_0 will seek to avoid damage, that R_0 knows that staying in the room will lead to its destruction, and so on. It is not a detailed model of R_0's circuits and memory-registers, still less is it a superbly detailed model that accounts for the movement of each molecule.

Can Robbie use R_0 as a self-model? After all, R_0 is in a perilous situation identical to Robbie's, and its knowledge and goals mirror Robbie's exactly. However, R_0 is not processing information in the same way that Robbie is. In particular, R_0 is not modeling another robot. For this reason let us reject R_0 as a possible self-model for Robbie. $R \neq R_0$.

So let Robbie use a more accurate model. Suppose Robbie models a robot R_1 that in turn models a robot R_0 such that R_0 perceives that a bomb is present, tentatively predicts its own demise, and exits the room (again, as above). Can Robbie use R_1 as a self-model? Like R_0, R_1 is in a perilous situation identical to Robbie's, and its knowledge and goals mirror Robbie's exactly. Unlike R_0, R_1 is modeling another robot. However the robot that R_1 is modeling is a simpler robot than Robbie is modeling. Like a Russian doll that contains another doll, Robbie's model contains a robot inside a robot,

while R_1's model is only 1 robot deep (it only includes the robot R_0). So we might reject R_1 as a possible self-model for Robbie. $R \neq R_1$.

Continue this process with R_2, R_3, R_4 and so on. None of them is quite accurate; when Robbie is modeling R_n , its model is n+1 robots deep, while the model of R_n is only n robots deep. Robbie is modeling R_n and the n robots inside of R_n, for a total of n+1 robots. R_n is only modeling its n robots. The only model that is completely accurate is R_∞, which requires infinite resources. So while technically it is impossible for Robbie to model itself with complete accuracy, it is possible to get very close. And anyway the differences between the models do not seem to be important; all the robots in all the models conclude that that they should exit the room. If Robbie chooses any one of them to serve as the self-model R, that model will accurately predict Robbie's action; will predict, that is, that Robbie will exit the room. The claim that Robbie cannot accurately model its own decision making process is false, even though some of the causal antecedents of its actions are situated in the causal-analysis box that is trying to analyze them.

4.5 An Alternative Argument

So McDermott's argument doesn't work. Is there another way to justify his conclusion? Let us return to McDermott's observation that "some of the causal antecedents of R's behavior are situated in the very causal-analysis box that is trying to analyze them." (p. 97) This rules out a plummeting robot situation (since then the analysis is causally inert), and adds

weight to the generous reading of condition 2 (that the selection is between meaningful actions). But it also seems to involve Robbie in a paradox: in order to predict what R is going to do, Robbie has to know what is going to happen to R, and to know that, Robbie first has to know what R is going to do. Although in some cases the paradox can be resolved (as by modeling R by a sequence of converging approximations) can we be sure that an analogous procedure will always work?

It would seem not. Suppose Robbie were confronted by Newcombe's problem. He is presented with two boxes, one opaque and one transparent. The transparent box contains $1000, the opaque box is either empty or else it contains $1 000 000. Robbie wants to maximize his take, and to do so can either take the contents of the opaque box (call this option A), or he can take the contents of both boxes (call this option B). There is a catch, though. A computer scientist has predicted Robbie's choice, and the contents of the opaque box have been determined accordingly. If the prediction is that Robbie will choose both boxes, then that box is empty. If the prediction is that Robbie will only take the opaque box, then that box contains $1 000 000. The computer scientist is very reliable (she has a duplicate of Robbie to put in the exact same situation), but the prediction and the disposition of the money have already been made and are irrevocable.

Let's see what happens if Robbie uses the method that worked so well in the bomb scenario. Robbie models what a simple robot like R_0 would do. R_0 (who, in the absence of a decision, can't account for what the computer

scientist might do) might reason as follows: the opaque box contains a non-negative quantity of money, a quantity we can call X. The transparent box contains $1000. So the choice is between X and X+1000. Clearly the latter is better; R_0 will choose B.

But consider what R_1 would conclude. R_1 simulates R_0, and notes that R_0 will choose B. But now R_1 has information relevant to what the computer scientist has done, namely with-hold the cash; R_1 predicts (based on the reliability of the computer scientist) that R_0 will find the opaque box empty, and will only get the $1000 in the transparent box. If R_1 confirms the decision to choose B, it will only get $1000. R_1 will then consider the result of revising the choice from B to A. R_1 models what happens if R_0 had chosen A, and discovers that a choice of A is perfectly correlated with $1 000 000 in the opaque box. So the choice, for R_1, is between a take of $1 000 000 and $1000. Clearly the former is better; R_1 will choose A.

Now what will R_2 do? If it assumes (implicitly or otherwise) that the computer scientist has based her prediction (and assignment of money) on the basis of what R_1 would do, then R_2 will infer that the opaque box is full of money. There is nothing preventing it from taking both boxes, and since two boxes of money are better than one box of money, it will choose B.

And so on. Even-numbered robots will choose B and odd-numbered robots will choose A. The approximations will not converge; what will Robbie use as a self-model?

To maximize his take Robbie has to be predicted to choose A but then choose B. Robbie cannot attempt to do this and at the same time have an accurate self-model, for as soon as he acts he falsifies the model he was basing his decision on. Now McDermott's conclusions seem to be true: under some circumstances a system finds that its decisions cannot be modeled in any available causal model, and therefore has to regard those decisions as undetermined. It must therefore regard itself as an agent who, in some special way, is exempt from causation.

Perhaps this is a faulty method of analyzing Newcombe's problem; Robbie, though smart, is not perfect. Nevertheless it illustrates McDermott's point. Anyway, besides Newcombe's problem there must be others situations where Robbie's capacities are insufficient for him to model R in a satisfactory manner. Our world is complex enough to guarantee that. The capacity to make these kinds of decisions (those which cannot be modeled in a system's self-model) would be used in other situations, too. There might be a situation where it requires too much time or resources to determine the optimal solution. Or there might be times when the value of being unpredictable (by competitors or predators, say) outweighs the marginal benefit of the otherwise optimal choice. In any event, systems which can make their way in a world like ours must have some way of making decisions even when the self-model fails. Let us modify McDermott's argument slightly:

If a system

1. Models its own behavior in terms of causes
2. Uses the output of the model to select among actions

3. Has beliefs about its own decisions

Then,

V. It will not always be able to model its own decision-making process

W. It will need a way of making decisions even when its model fails

X. It will believe that **at least some of** its own decisions are undetermined

Y. It will mark the symbols denoting itself as exempt from causality, and

Z. This is all that is meant by free will.

V is demonstrated by the Newcombe's problem example. I am reading into 2 that a failure of the model would not prevent an action from being selected; W just makes that reading explicit. The bomb example shows that the self-model need not always fail; I have modified X accordingly. X, Y and Z are basically an unpacking of the meaning of "cannot be causally modeled."

Recall that some terms used in this argument are used in a non-standard way; McDermott's use of "exempt from causality" and "undetermined" means that a system's decisions are not subject to either a unique or a probabilistic determination. McDermott's claim is that most theories of free will (libertarian as well as determinist) misdescribe our experience of decision making.

One key observation regarding W; the method the system uses cannot be one that it can model. This leads to a very interesting and rather poetic conclusion, namely that an agent must be a mystery to himself/herself in order to be free.

4.6 More about Free Will

This analysis of freedom requires that various situations be distinguished. The first situation corresponds to Robbie moving away from a bomb. Robbie is able to successfully model himself, since the series of approximate models all agree. The second situation corresponds to a Newcombe's problem kind of situation, where each approximation disagrees with its predecessor. Other possibilities could be imagined; the models might disagree at first, but eventually settle down and converge to a stable and optimal solution. Or perhaps the first approximations will agree, but later approximations might come to different solutions. Or perhaps successive approximations will agree with occasional dissenters; the longer the approximations are calculated, the more uncommon the dissenting models become. There are many possible variations.

A robot which has a "free will module" that allows it to choose from a non-convergent series of models must be able to cope with these different situations. The exact operation of the free will module cannot be modeled by the robot, but it is clear that the module must operate under certain constraints.[21] The module must allow the robot to use an appropriate amount of time to analyze its options; more time when there are no demands on the robot's attention, and less time when confronted by an urgent situation. The module must tend to choose options which are highly valued over ones that have a low utility. Options which are appraised as equally valuable should have roughly equal chances of being chosen. And so on.

[21] The "must" is the constraint of good design; what a savvy programmer (or the ruthless hand of natural selection) would put in.

A robot which considers a number of alternative future decisions and actions will view them as causally open; which alternative is chosen is up to the free will module, and its operation cannot be modeled. Looking back on a situation before the free will module acted, the robot will recognize that any of a number of options could have been chosen. Except for rare cases like the bomb in the room, the robot will not be able to model its choice as causally determined, for the choice was due to the free will module, and its operation cannot be modeled.

This situation bears an uncanny similarity to the "gap" described by Searle in his *Rationality in Action*.

> The gap can be given two equivalent descriptions, one forward-looking, one backward. Forward: the gap is that feature of our conscious decision making and acting where we sense alternative future decisions and actions as causally open to us. Backward: the gap is that feature of conscious decision making and acting whereby the reasons preceding the decisions and the action are not experienced by the agent as setting causally sufficient conditions for the decisions and actions. As far as our conscious experiences are concerned, the gap occurs when the beliefs, desires, and other reasons are not experienced as causally sufficient conditions for a decision... (*Rationality in Action*, 62)

The curious thing about this "gap" is that it has exists only from the perspective of Robbie's self-model. A simulation of Robbie by a more powerful computer need not reveal any gap, for the "free will module" might operate

deterministically- the only constraint is that the exact effect of the free will module's operation cannot be modeled by the robot it belongs to. When a more powerful computer models the free will module it could easily determine that at any point in time only one future was physically possible for Robbie. It is from the third-person perspective that Robbie's freedom is revealed as an illusion.

As far as a theory goes, McDermott's is obviously incomplete. There is no account of moral responsibility, of being aligned toward value, or of the ethics of reward and punishment. While his theory of qualia can address the various puzzles and issues in the literature, his theory of free will barely gets started.

On the other hand, this incompleteness makes McDermott's theory of free will compatible with any non-dualistic theory. It can easily be read as a hard determinist theory – it does, after all, make the claim that physical processes are (at least probabilistically) determined and that free will is an illusion. But it also distinguishes between situations of constraint (like being thrown from an airplane) and situations where the agent does what it wants (the exploding bomb case). This should please any compatibilists out there. Theorists who speculate about the way mental illness or defective socialization affect our freedom can make analogous speculations concerning the programming of the robot. Indeterminists can take heart from the indeterminacy of the "free will module." There is something for everyone.

Conclusion

McDermott's contribution to the philosophy of mind seems very modest. He took a few pages of Chalmers, concerning what the best reductionist theory of consciousness would be, and added the notion that the mind is a self-model. The result of this apparently minor insight is, as I have argued, a powerful theory that accounts for phenomenal consciousness and also makes a significant contribution to theories about free will. This is a remarkable accomplishment for a non-dualist and basically functionalist approach to the mind.

Although the notion of a simulation is implicit in any talk about computational models, and even though McDermott does refer to the illusory nature of mental phenomena, the elaboration of McDermott's theory with material on virtual reality goes well beyond anything McDermott discusses in his book. I admit this freely.

However, I hold the view that the primary task of philosophers is not to engage in careful exegesis of an author, but rather to plunder that author for ideas. Tell what the author should have said in addition to what the author did say. I have plundered McDermott (and Deutsch too) and I believe the theory I present as his is a charitable interpretation, what he should have said, the best and most convincing version I can put together of the theory presented in *Mind and Mechanism*.

Literature Cited

Block, N., "Troubles with Functionalism" in C.W. Savage, ed., *Perception and Cognition: Issues in the Foundations of Psychology* (Minneapolis: University of Minnesota Press, 1978). Reprinted in N. Block, ed., *Readings in the Philosophy of Psychology*, vol. 1. (Cambridge, MA: Harvard University Press, 1980)

Chalmers, David J., *The Conscious Mind* (Oxford: Oxford University Press, 1996)

_____, "Does a Rock Implement Every Finite-State Automaton?" *Synthese* 108 (1996) 309-333

_____, ed., *Philosophy of Mind* (Oxford: Oxford University Press, 2002)

Churchland, Paul, "Reduction, Qualia, and the Direct Introspection of Brain States" *Journal of Philosophy* 82 (1985) 8-28

Conee, Earl, "Phenomenal Knowledge" *Australasian Journal of Philosophy* 72 (1994) 136-150. Reprinted in Ludlow et al, eds., *There's Something About Mary*

Cooney, Brian, ed., *The Place of Mind* (Stanford, CT: Wadsworth, 2000)

Copeland, Jack, *Artificial Intelligence* (Oxford: Blackwell, 1993)

Dawkins, Richard, *The Selfish Gene* (Oxford: Oxford University Press, 1976, 2nd ed. 1989)

Dennett, Daniel, *Brainchildren* (Cambridge, MA: MIT Press, 1998)

Deutsch, David, *The Fabric of Reality* (New York: Penguin Books, 1997)

Domsky, Darren, *Is Syntax Intrinsic to Physics?* (Edmonton: University of Alberta MA Thesis, 1998)

Ludlow, Peter, Yujin Nagasawa and Daniel Stoljar eds., *There's Something About Mary* (Cambridge, MA: MIT Press, 2004)

McDermott, Drew V. *Mind and Mechanism* (Cambridge, MA: MIT Press, 2001)

Nida-Rümelin, Martine, "Pseudonormal Vision: An Actual Case of Qualia Inversion?" *Philosophical Studies* 82 (1996) 145-157. Reprinted in David Chalmers, ed., *Philosophy of Mind*

Nozick, Robert, *Anarchy, State, and Utopia* (New York: Basic Books, 1974)

Preston, John, ed., *Thought and Language* (Cambridge: Cambridge University Press, 1997)

Putnam, Hilary, *Representation and Reality* (Cambridge, MA: MIT Press, 1988)

Searle, J.R., "Minds, Brains, and Programs" *The Behavioral and Brain Sciences* 3 (1980) 417-419. Reprinted in Cooney, ed., *The Place of Mind*

_____, "Is the Brain a Digital Computer?" *Proceedings and Addresses of the American Philosophical Association* 64 (November 1990) 21-37. Reprinted in Cooney, ed., *The Place of Mind*

_____, *The Rediscovery of the Mind* (Cambridge, MA: MIT Press, 1992)

_____, "The Explanation of Cognition" in John Preston, ed., *Thought and Language*

_____, *Rationality in Action* (Cambridge, MA: MIT Press, 2001)

Wilson, Robert A., *Boundaries of the Mind* (Cambridge: Cambridge University Press, 2004)